Fast Proximal Gradient Descent for A Class of Non-convex and Non-smooth Sparse Learning Problems

Yingzhen Yang
School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
yingzhen.yang@asu.edu

Jiahui Yu
Beckman Institute
University of Illinois at Urbana-Champaign
jyu79@illinois.edu

Abstract

Non-convex and non-smooth optimization problems are important for statistics and machine learning. However, solving such problems is always challenging. In this paper, we propose fast proximal gradient descent based methods to solve a class of non-convex and non-smooth sparse learning problems, i.e. the ℓ^0 regularization problems. We prove improved convergence rate of proximal gradient descent on the ℓ^0 regularization problems, and propose two accelerated versions by support projection. The proposed accelerated proximal gradient descent methods by support projection have convergence rates which match the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continuous gradient. Experimental results demonstrate the effectiveness of the proposed algorithms. We also propose feed-forward neural networks as fast encoders to approximate the optimization results generated by the proposed accelerated algorithms.

1 INTRODUCTION

Non-convex and non-smooth optimization problems are challenging ones which have received a lot of attention in the machine learning literature (Bolte et al., 2014; Ochs et al., 2015). In this paper, we consider fast optimization algorithms for a class of non-convex and non-smooth sparse learning problems, i.e. the ℓ^0 regularized problems, presented as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}), \tag{1}$$

where $h(\mathbf{x}) \triangleq \lambda\|\mathbf{x}\|_0$, $\lambda > 0$ is a weighting parameter. g is assumed to satisfy the following two conditions throughout this paper:

(a) g is convex and g has bounded gradient, i.e. $\|\nabla g(\mathbf{x})\|_\infty \leq G$ for some constant G and any $\mathbf{x} \in \mathbb{R}^n$,

(b) g has L-Lipschitz continuous gradient, i.e. $\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$.

Assumption (b) is standard in the analysis of using proximal method for non-convex problems, e.g. (Bolte et al., 2014; Ghadimi and Lan, 2016a). When g is a G-Lipschitz continuous function, its gradient is bounded by G.

When g is the squared loss, i.e.

$$g(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2, \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{D}$ is the design matrix of dimension $d \times n$, (1) is the well-known ℓ^0 penalized Least Square Estimation (LSE) problem. Due to the nonconvexity imposed by the ℓ^0 regularization, extensive existing works resort to solve its ℓ^1 relaxation by linear programming or iterative shrinkage algorithms (Daubechies et al., 2004; Elad, 2006; Bredies and Lorenz, 2008; Agarwal et al., 2012). Albeit the nonconvexity of (1), sparse coding methods such as (Mancera and Portilla, 2006; Bao et al., 2014) that directly optimize virtually the same objective as (1) demonstrate compelling performance compared to its ℓ^1 norm counterparts in machine learning and computer vision. Cardinality constraint in terms of ℓ^0-norm is also studied for M-estimation problems by Iterative Hard-Thresholding (IHT) algorithm proposed by (Blumensath and Davies, 2008), e.g. (Jain et al., 2014; Shen and Li, 2017). Fast gradient based methods have been studied in the optimization literature (Nesterov, 2005, 2013; Tseng, 2008; Ghadimi and Lan, 2016b) for non-convex problems, and general iterative shrinkage and thresholding algorithm has been applied to problems with non-convex sparse regularization (Gong et al., 2013).

Due to the special property of ℓ^0 regularization function h, any critial point of g is also a critical point of F. However, finding a sparse critical point of F is still challenging

due to the non-convexity and nonsmoothness of h. We use Proximal Gradient Descent (PGD) method to obtain a sparse sub-optimal solution to (1), and proposed fast PGD methods with guarantee on fast convergence rate by a novel operation named support shrinkage. Our main results are summarized in the subsection below.

1.1 Main Results

We extend the existing Accelerated Proximal Gradient descent method (APG) (Beck and Teboulle, 2009a,b) on the ℓ^0 regularization problem (1), showing improved convergence rates. More concretely,

- We present Theorem 2 showing that the sequence generated by PGD converges to a critical point of F, with convergence rate

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \leq \mathcal{O}(\frac{1}{m}) \tag{3}$$

  for all $m \geq k_0$ with some $k_0 \geq 0$.

- We propose two new accelerated versions of PGD with a novel support projection operation, namely Nonmonotone Accelerated Proximal Gradient Descent with Support Projection (NAPGD-SP) and Monotone Accelerated Proximal Gradient Descent with Support Projection (MAPGD-SP). While facing highly non-convex and non-smooth problem (1), both algorithms match the Nesterov's optimal convergence rate of first-order methods on smooth and convex objective function with Lipschitz continuous gradient. For both NAPGD-SP and MAPGD-SP, we show the following convergence rates:

$$\|F(\mathbf{x}^{(m)}) - F(\mathbf{x}^*)\|_2 \leq \mathcal{O}(\frac{1}{m^2}), \tag{4}$$

  for all $m \geq k_0$ with some $k_0 \geq 0$. Please refer to Theorem 3 and Theorem 4 in Section 4 for details. It should be emphasized that this is the same convergence rate as that of regular APG on convex problems (Beck and Teboulle, 2009a,b)

The obtained sub-optimal solutions by PGD, NAPGD-SP and MAPGD-SP are all sparser than the initialization point, and the support of each obtained solution is a subset of that of the initialization, which enables interpretable variable shrinkage.

The general accelerated algorithms in (Li and Lin, 2015) have linear convergence rate with $\theta \in [1/2, 1)$, and sublinear rate $\mathcal{O}(k^{-\frac{1}{1-2\theta}})$ with $\theta \in (0, 1/2)$, where θ is the Kurdyka- Lojasiewicz (KL) exponent and both rates are for objective values. To the best of our knowledge, the machine learning and optimization literature has no concrete results on which interval θ lies in ($(0, 1/2)$ or $[1/2, 1)$), when the sequence is approaching to a critical point of problem (1) using general algorithms such as those in (Li and Lin, 2015) and (Bolte et al., 2014). θ could take values in both intervals as the sequence approaches to a critical point. Therefore, one can only claim a conservative sublinear convergence rate ($\mathcal{O}(k^{-\frac{1}{1-2\theta}})$ and $\theta \in (0, 1/2)$) for problem (1) (e.g., by (Bao et al., 2014)). The algorithms in our work impose support projection so that the sequence has the same support after finite k_0 iterations. By the support shrinkage property, we prove the convergence rate $\mathcal{O}(\frac{1}{k})$ for vanilla PGD and $\mathcal{O}(\frac{1}{k^2})$ for the proposed NAPGD-SP and MAPGD-SP on problem (1), which are not guaranteed by the general algorithms in (Li and Lin, 2015). It can also be verified that they achieve linear convergence rates when g is strongly convex. Again, the convergence results mentioned above and presented in this paper can not be guaranteed by the general algorithms in (Li and Lin, 2015) with analysis via KL property, to the best of our knowledge. Our results are among the very few results in the machine learning and optimization literature about fast optimization methods on non-convex and non-smooth problems involving ℓ^0 regularization. Moreover, our methods reveal the hidden convexity in ℓ^0 regularization problems and have the potential to be extended to more general non-convex and non-smooth problems.

1.2 Notations

Throughout this paper, we use bold letters for matrices and vectors, regular lower letters for scalars. The bold letter with subscript indicates the corresponding element of a matrix or vector, and the bold letter with superscript indicates the corresponding column of a matrix, i.e. $\mathbf{D}^i$ indicates the i-th column of matrix $\mathbf{D}$. $\|\cdot\|_p$ denotes the ℓ^p-norm of a vector, or the p-norm of a matrix. We let $\boldsymbol{\beta}_{\mathbf{I}}$ denote the vector formed by the elements of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a vector, or matrix formed by columns of $\boldsymbol{\beta}$ with indices indices in $\mathbf{I}$ when $\boldsymbol{\beta}$ is a matrix. $\mathrm{supp}(\cdot)$ indicates the support of a vector, i.e. the set of indices of nonzero elements of this vector. $\sigma_t(\cdot)$ is the t-th largest singular value of a matrix, and $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ indicate the smallest and largest singular value of a matrix respectively. $|\mathbf{A}|$ denotes the cardinality of a set $\mathbf{A}$, $\mathrm{supp}(\cdot)$ indicates the support of a vector. We denote by $\mathrm{Col}(\mathbf{A})$ the subspace spanned by columns of matrix $\mathbf{A}$. We let $\mathbf{S} = \mathrm{supp}(\mathbf{x}^{(0)})$ be the support of the initialization point $\mathbf{x}^{(0)}$ for optimization, and $D = \max_i \|\mathbf{D}^i\|_2$.

2 ALGORITHMS

We introduce PGD and their accelerated versions with support projection in this section.

2.1 Proximal Gradient Descent for ℓ^0 Sparse Approximation

The ℓ^0 regularization problem (1) is NP-hard in general (Natarajan, 1995). Therefore, the literature extensively resorts to approximate algorithms, such as Orthogonal Matching Pursuit (Tropp, 2004), or that using surrogate functions (Hyder and Mahata, 2009). In addition, PGD has been used by (Bao et al., 2014) to find an approximate solution to (1) with g being the squared loss (2), and that method is proved to have sublinear convergence rate with satisfactory empirical results. The success of PGD raises an interesting question that what is the convergence rate of PGD on the general ℓ^0 regularization problem (1).

In this section, we first present PGD to optimize (1) in an iterative shrinkage manner. Then we introduce two of its accelerated versions, namely Nonmonotone Accelerated Proximal Gradient Descent with Support Projection (NAPGD-SP) and Monotone Accelerated Proximal Gradient Descent with Support Projection (MAPGD-SP). Both NAPGD-SP and MAPGD-SP feature support projection for fast convergence.

2.2 Proximal Gradient Descent

In the k-th iteration of PGD for $k \geq 1$, gradient descent is performed on the squared loss term $g(\mathbf{x})$ to obtain an intermediate variable as the result of gradient descent, i.e. $\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})$, where $s > 0$ is the step size, and $\frac{1}{s}$ is usually chosen to be larger than a Lipschitz constant L for the gradient of function $g(\cdot)$, namely $\|\nabla g(\mathbf{u}) - \nabla g(\mathbf{v})\|_2 \leq L\|\mathbf{u} - \mathbf{v}\|_2$ for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$.

The proximal mapping associated with h is defined as $\text{prox}_h(\mathbf{u}) = \arg\min_{\mathbf{v}} h(\mathbf{v}) + \frac{1}{2}\|\mathbf{v} - \mathbf{u}\|_2^2$. $\mathbf{x}^{(k+1)}$ is then the solution to the following the proximal mapping on $\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})$:

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \text{prox}_{sh}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})) \\ &= \arg\min_{\mathbf{v}\in\mathbb{R}^n} \frac{1}{2s}\|\mathbf{v} - (\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))\|_2^2 + \lambda\|\mathbf{v}\|_0 \\ &= T_{\sqrt{2\lambda s}}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})), \end{aligned} \tag{5}$$

where T_θ is an element-wise hard thresholding operator:

$$[T_\theta(\mathbf{u})]_j = \begin{cases} 0 & : \ |\mathbf{u}_j| \leq \theta \\ \mathbf{u}_j & : \ \text{otherwise} \end{cases}, \quad 1 \leq j \leq n.$$

The iterations start from $k = 1$ and continue until the sequence $\{F(\mathbf{x}^{(k)})\}_k$ or $\{\mathbf{x}^{(k)}\}_k$ converges or maximum iteration number is achieved. The optimization algorithm for the ℓ^0 sparse approximation problem (1) by PGD is described in Algorithm 1. In practice, the time complexity of optimization by PGD is $\mathcal{O}(Mdn)$ where M is the number of iterations (or maximum number of iterations) for PGD.

Algorithm 1 Proximal Gradient Descent for the ℓ^0 Regularization Problem (1)

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$.

1: **for** $k = 0, \ldots,$ **do**
2: Update $\mathbf{x}^{(k+1)}$ according to (5)
3: **end for**

Output: Obtain the sparse solution $\hat{\mathbf{x}}$ upon the termination of the iterations.

2.3 Accelerated Proximal Gradient Descent with Support Projection

The Nonmonotone Accelerated Proximal Gradient Descent with Support Projection (NAPGD-SP) and Monotone Accelerated Proximal Gradient Descent (MAPGD-SP) with Support Projection (MAPGD-SP) are introduced in the following two subsections. Both algorithms are feministic of regular APG (Beck and Teboulle, 2009a,b), and they achieve fast convergence by virtue of support projection to be explained in the next subsections.

2.3.1 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection

The update rules in the k-th iteration of NAPGD-SP are presented as follows.

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k)} + \frac{t_{k-1} - 1}{t_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \tag{6}$$

$$\mathbf{w}^{(k)} = \mathbf{P}_{\text{supp}(\mathbf{x}^{(k)})}(\mathbf{u}^{(k)}), \tag{7}$$

$$\mathbf{x}^{(k+1)} = \text{prox}_{sh}(\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)})), \tag{8}$$

$$t_{k+1} = \frac{\sqrt{1 + 4t_k^2} + 1}{2}, \tag{9}$$

where $\mathbf{P}_{\mathbf{S}'}(\mathbf{u})$ indicates the novel support projection operator which returns a vector whose elements with indices in $\mathbf{S}'$ are the same as those in $\mathbf{u}$, while all the other elements vanish. Note that (6), (8) and (9) also appear in regular nonmonotone APG (Beck and Teboulle, 2009b) for convex problems, and support projection is employed to enforce the support shrinkage property so as to guarantee fast convergence for the non-convex ℓ^0 regularization problem. The algorithm for NAPGD-SP is shown in Algorithm 2.

We say that a strict support shrinkage happens if $\text{supp}(\mathbf{x}^{(k+1)}) \subset \text{supp}(\mathbf{x}^{(k)})$. Algorithm 2 describes NAPGD-SP for the ℓ^0 regularization problem (1).

2.3.2 Monotone Accelerated Proximal Gradient Descent

The update rules in the k-th iteration of MAPGD-SP are presented as follows. The algorithm for MAPGD-SP is

Algorithm 2 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection for the ℓ^0 Regularization Problem (1)

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$, $\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$.

1: **for** $k = 1, \ldots,$ **do**
2: Update $\mathbf{u}^{(k)}$, $\mathbf{w}^{(k)}$, $\mathbf{x}^{(k+1)}$, t_{k+1} according to (6), (7), (8), (9) respectively.
3: **end for**

Output: Obtain the sparse solution $\hat{\mathbf{x}}$ upon the termination of the iterations.

shown in Algorithm 3. Again, (10), (12) and (13) also in regular monotone APG (Beck and Teboulle, 2009a) for convex problems, and support projection is introduced in (11) to constrain the support of the intermediate variable $\mathbf{w}^{(k)}$.

$$\mathbf{u}^{(k)} = \mathbf{x}^{(k)} + \frac{t_{k-1}}{t_k}(\mathbf{z}^{(k)} - \mathbf{x}^{(k)}) + \frac{t_{k-1}-1}{t_k}(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}), \tag{10}$$

$$\mathbf{w}^{(k)} = \mathbf{P}_{\mathrm{supp}(\mathbf{z}^{(k)})}(\mathbf{u}^{(k)}), \tag{11}$$

$$\mathbf{z}^{(k+1)} = \mathrm{prox}_{sh}(\mathbf{w}^{(k)} - s\nabla g(\mathbf{w}^{(k)})), \tag{12}$$

$$t_{k+1} = \frac{\sqrt{1+4t_k^2}+1}{2}, \tag{13}$$

$$\mathbf{x}^{(k+1)} = \begin{cases} \mathbf{z}^{(k+1)} & \text{if } F(\mathbf{z}^{(k+1)}) \le F(\mathbf{x}^{(k)}) \\ \mathbf{x}^{(k)} & \text{otherwise.} \end{cases} \tag{14}$$

Algorithm 3 Monotone Accelerated Proximal Gradient Descent with Support Projection for the ℓ^0 Regularization Problem (1)

Input:
The weighting parameter λ, the initialization $\mathbf{x}^{(0)}$, $\mathbf{z}^{(1)} = \mathbf{x}^{(1)} = \mathbf{x}^{(0)}$, $t_0 = 0$.

1: **for** $k = 1, \ldots,$ **do**
2: Update $\mathbf{u}^{(k)}$, $\mathbf{w}^{(k)}$, $\mathbf{z}^{(k+1)}$, t_{k+1}, $\mathbf{x}^{(k+1)}$ according to (10), (11), (12), (13), and (14) respectively.
3: **end for**

Output: Obtain the sparse solution $\hat{\mathbf{x}}$ upon the termination of the iterations.

3 ANALYSIS OF PROXIMAL GRADIENT DESCENT

In this section we present the analysis for the convergence rate of PGD in Algorithm 1. We first present the support shrinkage property in the following lemma, showing that the support of the sequence $\{\mathbf{x}^{(k)}\}_k$ shrinks.

Lemma 1. (Support shrinkage for proximal gradient descent in Algorithm 1 and sufficient decrease of the objective function) *If* $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, *then*

$$\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), k \ge 0, \tag{15}$$

namely the support of the sequence $\{\mathbf{x}^{(k)}\}_k$ *shrinks. Moreover, the sequence of the objective* $\{F(\mathbf{x}^{(k)})\}_k$ *is nonincreasing, and the following inequality holds for* $k \ge 0$*:*

$$F(\mathbf{x}^{(k+1)}) \le F(\mathbf{x}^{(k)}) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2. \tag{16}$$

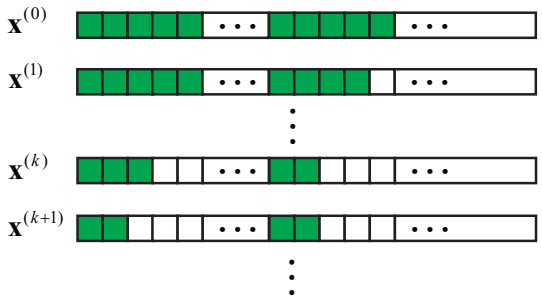


Figure 1: Illustration of the support shrinkage property shown in Lemma 1. Each green box indicates a nonzero element, and each white box indicate a zero element. As the iteration proceeds, some of the green boxes gradually turn white while no white box turn green, which reflects the support shrinkage property, i.e. $\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)})$ for $k \ge 0$.

Figure 1 illustrates the support shrinkage property. Under the mild conditions in Lemma 1, the support shrinkage property (15) holds, and $|\mathrm{supp}(\mathbf{x}^{(k)})| \le |\mathrm{supp}(\mathbf{x}^{(k)})|$. Given fixed λ, one may be concerned that a very small step for the gradient descent is required to ensure the support shrinkage property. Note that when g is chosen as (2) for sparse approximation problem, the canonical choice for L is $2\|\mathbf{D}\|_2^2$. Let $\|\mathbf{y} - \mathbf{D}\mathbf{x}^{(0)}\|_2^2 = x_0$, and it can be verified that $G \le 2D\sqrt{x_0 + \lambda|\mathbf{S}|}$. If the size of the dictionary n is moderately large compared to d and $|\mathbf{S}|$, then we show in Theorem 1 that $\frac{2\lambda}{G^2} \ge \frac{1}{L}$, i.e. our choice for s in Lemma 1 would not lead to a smaller step size for gradient descent with high probability, compared to the case when the conventional choice (with s less than $\frac{1}{L}$) is adopted.

Theorem 1. *Suppose* $\mathbf{D} \in \mathbb{R}^{d\times n}$ $(n \ge d)$ *is a random matrix whose elements are i.i.d. samples from the standard Gaussian distribution* $\mathcal{N}(0,1)$. *Then with probability at least* $1 - e^{-\frac{nt^2}{2}} - ne^{-t}$,

$$\frac{2\lambda}{G^2} \ge \frac{1}{L} \tag{17}$$

if

$$n \geq \big(\sqrt{d} + t + \sqrt{\frac{(d + 2\sqrt{dt} + 2t)(x_0 + \lambda|\mathbf{S}|)}{\lambda}}\big)^2, \tag{18}$$

and t can be chosen as $t_0 \log n$ for $t_0 > 0$ to ensure that (18) holds and (17) holds with high probability.

It follows from (15) that $0 \leq |\mathrm{supp}(\mathbf{x}^{(k)})| \leq |\mathbf{S}|$ for any $k \geq 0$, and the sequence $\{\mathbf{x}^{(k)}\}_k$ generated by Algorithm 1 according to (5) can be segmented into the following $|\mathbf{S}| + 1$ subsequences $\{\mathcal{X}^{(t')}\}_{t'=0}^{|\mathbf{S}|}$ with

$$\mathcal{X}^{(t')} = \{\mathbf{x}^{(k)} \colon |\mathrm{supp}(\mathbf{x}^{(k)})| = t', k \geq 0\}, \quad 0 \leq t' \leq |\mathbf{S}|. \tag{19}$$

It can be verified that $\bigcup_{t'=0}^{|\mathbf{S}|} \mathcal{X}^{(t')} = \{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$, and $\mathcal{X}^{(t_1)} \bigcap \mathcal{X}^{(t')} = \emptyset$. Therefore, $\{\mathcal{X}^{(t')}\}_{t'=0}^{|\mathbf{S}|}$ forms a disjoint cover of the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$. We are certainly interested in the nonempty subsequences in $\{\mathcal{X}^{(t')}\}_{t'=0}^{|\mathbf{S}|}$, which are formally defined as subsequences with shrinking support in Definition 1. These nonempty subsequences still form a disjoint cover of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ and they are in descending order of the support size.

Definition 1. (Subsequences with shrinking support) *All the $T \leq |\mathbf{S}| + 1$ nonempty subsequences among $\{\mathcal{X}^{(t')}\}_{t'=0}^{|\mathbf{S}|}$ are defined to be subsequences with shrinking support, denoted by $\{\mathcal{X}_t\}_{t=1}^{T}$. The subsequences with shrinking support form a disjoint are ordered with decreasing support size, i.e. $|\mathrm{supp}(\mathbf{x}^{(k_2)})| < |\mathrm{supp}(\mathbf{x}^{(k_1)})|$ for any $\mathbf{x}^{(k_1)} \in \mathcal{X}_{t_1}$ and $\mathbf{x}^{(k_2)} \in \mathcal{X}_{t_2}$ with any $1 \leq t_1 < t_2 \leq T$.*

We have the following lemma about the properties of subsequences with shrinking support.

Lemma 2. (Properties of the subsequences with shrinking support)

(i) All the elements of each subsequence $\mathcal{X}_t$ ($t = 1, \ldots, T$) in the subsequences with shrinking support have the same support. In addition, for any $1 \leq t_1 < t_2 \leq T$ and any $\mathbf{x}^{(k_1)} \in \mathcal{X}_{t_1}$ and $\mathbf{x}^{(k_2)} \in \mathcal{X}_{t_2}$, we have $k_1 < k_2$, $\mathrm{supp}(\mathbf{x}^{(k_2)}) \subset \mathrm{supp}(\mathbf{x}^{(k_1)})$.

(ii) All the subsequence except for the last one, namely $\mathcal{X}_t$ ($t = 1, \ldots, T-1$), have finite size. Moreover, $\mathcal{X}_T$ has infinite number of elements, and there exists $k_0 \geq 0$ such that $\{\mathbf{x}^{(k)}\}_{k=k_0}^{\infty} \subseteq \mathcal{X}_T$.

Before stating Theorem 2 about the convergence rate of PGD, the definition of critical point is introduced as follows.

Definition 2. (Subdifferential and critical points) *Given a non-convex function $f \colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ which is a proper and lower semi-continuous function.*

- *for a given $\mathbf{x} \in \mathrm{dom} f$, its Frechet subdifferential of f at $\mathbf{x}$, denoted by $\tilde{\partial} f(\mathbf{x})$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^n$ which satisfy*

$$\liminf_{\mathbf{y} \neq \mathbf{x}, \mathbf{y} \to \mathbf{x}} \frac{f(\mathbf{y}) - f(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|} \geq 0.$$

- *The limiting-subdifferential of f at $\mathbf{x} \in \mathbb{R}^n$, denoted by written $\partial f(\mathbf{x})$, is defined by*

$$\partial f(\mathbf{x}) = \{\mathbf{u} \in \mathbb{R}^n \colon \exists \mathbf{x}^k \to \mathbf{x}, f(\mathbf{x}^k) \to f(\mathbf{x}), \tilde{\mathbf{u}}^k \in \tilde{\partial} f(\mathbf{x}^k) \to \mathbf{u}\}.$$

The point $\mathbf{x}$ is a critical point of f if $\mathbf{0} \in \partial f(\mathbf{x})$.

Denote by $\mathbf{S}^*$ the support of any element in $\mathcal{X}_T$. If $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1 has a limit point $\mathbf{x}^*$, then the following theorem shows that the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ converges to $\mathbf{x}^*$, and $\mathbf{x}^*$ is a critical point of $F(\cdot)$ whose support is $\mathbf{S}^*$.

Theorem 2. (Convergence of PGD for the ℓ^0 regularization problem (1)) *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, and $\mathbf{x}^*$ is a limit point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$. Then the sequence $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1 converges to $\mathbf{x}^*$, and $\mathbf{x}^*$ is a critical point of $F(\cdot)$. Moreover, there exists $k_0 \geq 0$ such that for all $m \geq k_0$,*

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \leq \frac{1}{2s(m - k_0 + 1)} \|\mathbf{x}^{(k_0)} - \mathbf{x}^*\|_2^2. \tag{20}$$

Remark 1. *It should be emphasize that k_0 is a bounded by a constant determined by $\mathbf{x}^0$,g, λ and $\mathbf{S}$. Please refer to the proof of this theorem in the supplementary document of this paper.*

4 ANALYSIS OF ACCELERATED PROXIMAL GRADIENT DESCENT WITH SUPPORT PROJECTION

We analyze the convergence rates of NAPGD-SP and MAPGD-SP in the following two subsections.

4.1 Nonmonotone Accelerated Proximal Gradient Descent with Support Projection

Lemma 3 shows the support shrinkage property for NAPGD-SP, based on which the convergence rate of NAPGD-SP is presented in Theorem 3.

Lemma 3. (Support shrinkage for nonmonotone accelerated proximal gradient descent with support projection in Algorithm 2) *The sequence $\{\mathbf{x}^{(k)}\}_k$ generated by Algorithm 2 satisfies*

$$\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), k \geq 1, \tag{21}$$

namely the support of the sequence $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ shrinks.

Denote by $\mathbf{S}^*$ the support of any element in $\mathcal{X}_T$. If $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 1 has a limit point $\mathbf{x}^*$, then the following theorem shows that the sequence $\{F(\mathbf{x}^{(k)})\}_{k=0}^{\infty}$ converges to $F(\mathbf{x}^*)$ with Nesterov's optimal convergence rate.

Theorem 3. (Convergence of Nonmonotone Accelerated Proximal Gradient Descent for the ℓ^0 regularization problem (1)) *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, and $\mathbf{x}^*$ is a limit point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 2. There exists $k_0 \geq 1$ such that*

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \leq \frac{4}{(m+1)^2} V^{(k_0)} \tag{22}$$

for all $m \geq k_0$, where

$$V^{(k_0)} \triangleq \Big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{x}^{(k_0)} + \mathbf{x}^*\|_2^2 + t_{k_0-1}^2\big(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*)\big)\Big). \tag{23}$$

4.2 Monotone Accelerated Proximal Gradient Descent with Support Projection

Lemma 4 shows the support shrinkage property for MAPGD-SP, based on which the convergence rate of MAPGD-SP is presented in Theorem 4.

Lemma 4. (Support shrinkage for accelerated proximal gradient descent with support projection in Algorithm 3) *The sequence $\{\mathbf{z}^{(k)}\}_{k=1}^{\infty}$ and $\{\mathbf{x}^{(k)}\}_{k=1}^{\infty}$ generated by Algorithm 3 satisfy*

$$\mathrm{supp}(\mathbf{z}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{z}^{(k)}), \tag{24}$$

$$\mathrm{supp}(\mathbf{x}^{(k+1)}) \subseteq \mathrm{supp}(\mathbf{x}^{(k)}), \tag{25}$$

namely the support of both sequences shrinks.

Similar to the nonmonotone case, denote by $\mathbf{S}^*$ the support of any element in $\mathcal{X}_T$. If $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 3 has a limit point $\mathbf{x}^*$, then the following theorem shows that the sequence $\{F(\mathbf{x}^{(k)})\}_{k=0}^{\infty}$ converges to $F(\mathbf{x}^*)$ with Nesterov's optimal convergence rate.

Theorem 4. (Convergence of Monotone Accelerated Proximal Gradient Descent for the ℓ^0 regularization problem (1)) *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, and $\mathbf{x}^*$ is a limit point of $\{\mathbf{x}^{(k)}\}_{k=0}^{\infty}$ generated by Algorithm 3. There exists $k_0 \geq 1$ such that*

$$F(\mathbf{x}^{(m+1)}) - F(\mathbf{x}^*) \leq \frac{4}{(m+1)^2} W^{(k_0)} \tag{26}$$

for all $m \geq k_0$, where

$$W^{(k_0)} \triangleq \big(\frac{1}{2s}\|(t_{k_0-1}-1)\mathbf{x}^{(k_0-1)} - t_{k_0-1}\mathbf{z}^{(k_0)} + \mathbf{x}^*\|_2^2 + t_{k_0-1}^2(F(\mathbf{x}^{(k_0)}) - F(\mathbf{x}^*))\big). \tag{27}$$

5 ROADMAP OF PROOFS

The key point in the proof of Theorem 2 is that, during the last stage of the optimization wherein the variable $\{\mathbf{x}^{(k)}\}_{k=k_0}^{\infty}$ has the same support $\mathbf{S}^*$, the optimization for the ℓ^0 regularization problem behaves the same as convex optimization. The same idea is employed to prove Theorem 3 and Theorem 4. In fact, $\mathbf{x}^{(k+1)}$ is the solution to the proximal mapping (5), so we have

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathrm{prox}_{sh}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)})) \\ &= \arg\min_{\mathbf{v}\in\mathbb{R}^n} \frac{1}{2s}\|\mathbf{v} - (\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))\|_2^2 + h(\mathbf{v}).\end{aligned} \tag{28}$$

It follows from (28) that

$$\begin{aligned}&\frac{1}{s}(\mathbf{z}^{(k+1)} - (\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))) + \partial h(\mathbf{x}^{(k+1)}) = 0 \\ &\Rightarrow -\nabla g(\mathbf{x}^{(k)}) - \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \in \partial h(\mathbf{x}^{(k+1)}).\end{aligned} \tag{29}$$

Since $\mathbf{x}^{(k+1)} = T_{\sqrt{2\lambda s}}(\mathbf{x}^{(k)} - s\nabla g(\mathbf{x}^{(k)}))$, we have $[\partial h(\mathbf{x}^{(k+1)})]_j = 0$ for any $j \in \mathrm{supp}(\mathbf{x}^{(k+1)})$. It follows that for any vector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\mathbf{x}^{(k+1)})$, the following equality holds:

$$\begin{aligned}h(\mathbf{v}) = h(\mathbf{x}^{(k+1)}) + \langle -\nabla g(\mathbf{x}^{(k)}) - \frac{1}{s}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}), \\ \mathbf{v} - \mathbf{x}^{(k+1)}\rangle.\end{aligned} \tag{30}$$

When h is convex and differentiable, then we have

$$h(\mathbf{v}) \geq h(\mathbf{x}^{(k+1)}) + \langle \nabla h(\mathbf{x}^{(k+1)}), \mathbf{v} - \mathbf{x}^{(k+1)}\rangle. \tag{31}$$

The inequality (31) is crucial in the proof of convergence guarantee for proximal gradient descent on convex problems. It is not guaranteed to hold when h is non-convex and non-smooth, e.g. when $h(\cdot) = \lambda\|\cdot\|_0$. However, due to the properties of thec and the solution to the proximal mapping associated with the ℓ^0-norm (28), (30), which has a similar form to (31), holds when $\mathbf{v}$ has the same support as that of $\mathbf{x}^{(k+1)}$. Therefore, in each stage of the optimization by Algorithm 1, the variable sequence has fixed support and (31) is applicable, leading to the convergence of proximal gradient descent on the nonconvex and non-smooth ℓ^0 regularization problem with the same convergence rate as that for the canonical convex problems.

To elaborate more details in proof of Theorem 2, due to the convexity and Lipschitz continuity of g we have

$$F(\mathbf{x}^{(k+1)}) = g(\mathbf{x}^{(k+1)}) + h(\mathbf{x}^{(k+1)})$$

$$\leq g(\mathbf{v}) + \langle \nabla g(\mathbf{x}^{(k)}), \mathbf{x}^{(k+1)} - \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 + h(\mathbf{x}^{(k+1)}). \quad (32)$$

According to Lemma 1, $\{F(\mathbf{x}^{(k)})\}_k$ is nonincreasing, it follows that $\{\mathbf{x}^{(k)}\}_k$ is a bounded sequence with a converging subsequence. It can be proved that the subsequence converges to a critical point of $F(\cdot)$, denoted by $\mathbf{x}^*$, and $\mathrm{supp}(\mathbf{x}^*) = \mathbf{S}^*$.

Now $\mathrm{supp}(\mathbf{x}^*) = \mathrm{supp}(\mathbf{x}^{(k+1)}) = \mathbf{S}^*$, we let $\mathbf{v} = \mathbf{x}^*$ and combine (30) and (32), we have

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq \frac{1}{s}\langle \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}, \mathbf{x}^* - \mathbf{x}^{(k)} \rangle - \frac{1}{2s}\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2$$
$$= \frac{1}{2s}(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|_2^2). \quad (33)$$

Summing (33) over $k = k_0, \ldots, m$ with $m \geq k_0$, we have

$$\sum_{k=k_0}^{m} F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^*) \leq \sum_{k=k_0}^{m} \frac{1}{2s}(\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\|_2^2)$$
$$= \frac{1}{2s}(\|\mathbf{x}^{(k_0)} - \mathbf{x}^*\|_2^2 - \|\mathbf{x}^{(m+1)} - \mathbf{x}^*\|_2^2), \quad (34)$$

which leads to the main conclusions Theorem 2.

6 DISCUSSION ABOUT CONSISTENCY OF THE OPTIMIZATION RESULTS

Another interesting question is that whether the critical point, $\mathbf{x}^*$, obtained by our proposed algorithms is optimal from the perspective of statistics. We consider the ℓ^0 penalized LSE problem presented as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^n} F(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + h(\mathbf{x}), \quad (35)$$

where $g(\mathbf{x}) = \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2$, $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{D}$ is the design matrix of dimension $d \times n$. Let $\bar{\mathbf{x}}^*$ be the globally optimal solution to (35), $\bar{\mathbf{S}}^* = \mathrm{supp}(\bar{\mathbf{x}}^*)$. The following theorem presents the bound between $\mathbf{x}^*$ and $\bar{\mathbf{x}}^*$.

Theorem 5. (Sub-optimal solution is close to the globally optimal solution for ℓ^0 penalized LSE) *Suppose $\mathbf{D}_{\mathbf{S}\cup\bar{\mathbf{S}}^*}$ has full column rank with $\kappa_0 \triangleq \sigma_{\min}(\mathbf{D}_{\mathbf{S}\cup\bar{\mathbf{S}}^*}) > 0$. Let $\kappa > 0$ such that $2\kappa_0^2 > \kappa$ and b is chosen according to (36) as below:*

$$0 < b < \min\{\min_{j\in\mathbf{S}^*} |\mathbf{x}_j^*|, \frac{\lambda}{\max_{j\notin\mathbf{S}^*} |\frac{\partial g}{\partial \mathbf{x}_j}|_{\mathbf{x}=\mathbf{x}^*}|}, \min_{j\in\bar{\mathbf{S}}^*} |\bar{\mathbf{x}}_j^*|, \frac{\lambda}{\max_{j\notin\bar{\mathbf{S}}^*} |\frac{\partial g}{\partial \mathbf{x}_j}|_{\mathbf{x}=\bar{\mathbf{x}}^*|}}\}. \quad (36)$$

Let $\mathbf{F} = (\mathbf{S}^* \setminus \bar{\mathbf{S}}^*) \cup (\bar{\mathbf{S}}^* \setminus \mathbf{S}^*)$ be the symmetric difference between $\mathbf{S}^*$ and $\bar{\mathbf{S}}^*$, then

$$\|\mathbf{x}^* - \bar{\mathbf{x}}^*\|_2 \leq \frac{1}{2\kappa_0^2 - \kappa}\Big(\sum_{j\in\mathbf{F}\cap\mathbf{S}^*} (\max\{0, \frac{\lambda}{b} - \kappa|\mathbf{x}_j^* - b|\})^2 + \sum_{j\in\mathbf{F}\setminus\mathbf{S}^*} (\max\{0, \frac{\lambda}{b} - \kappa b\})^2\Big)^{\frac{1}{2}}. \quad (37)$$

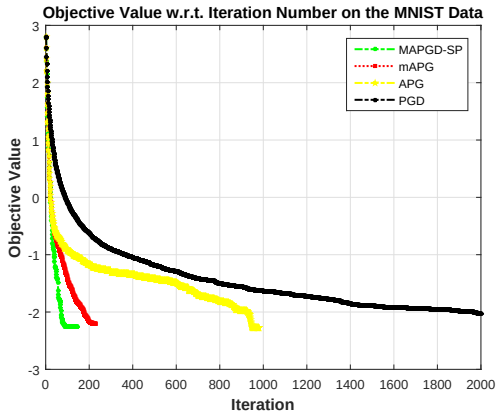
According to Theorem 5, when $\frac{\lambda}{b} - \kappa|\mathbf{x}_j^* - b|$ for nonzero $\mathbf{x}_j^*$ and $\frac{\lambda}{b} - \kappa b$ are no greater than 0, the sub-optimal solution $\mathbf{x}^*$ is equal to the globally optimal solution $\bar{\mathbf{x}}^*$. In this case, the support identification is guaranteed by the existing results in the support recovery property of ℓ^0 optimization in Theorem 4 of (Zhang and Zhang, 2012).

7 EXPERIMENTAL RESULTS

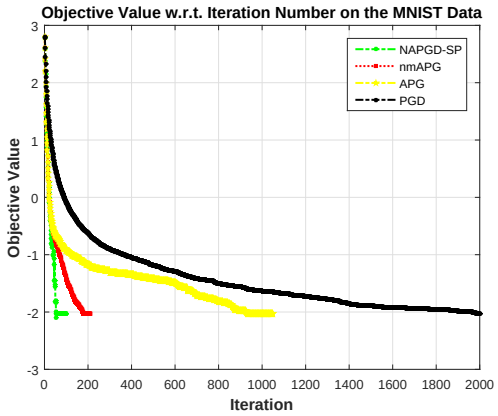
We demonstrate empirical results of the proposed fast PGD methods in this section. We first study the ℓ^0 penalized LSE problem where g is chosen as (2), and the initialization point guarantees that $\mathbf{D}_{\mathbf{S}}$ has full column rank. We conduct experiments on the MNIST handwritten digits database. The MNIST data set contains $70,000$ examples of handwritten digits, of which $10,000$ constitute a test set. Each example is a 28×28 grayscale image and represented by a 784-dimensional vector. We first run online dictionary learning (Mairal et al., 2010) to learn a dictionary matrix $\mathbf{D}$ of 300 columns. We then randomly choose an image as $\mathbf{y}$, and optimize problem (1) by PGD, NAPGD-SP and MAPGD-SP to obtain $\mathbf{x}^*$. We plot the illustration of the sequence of the objective values, i.e. $f(\mathbf{x}^{(k)})$, with respect to the iteration number k for monotone and nonmonotone algorithms in Figure 2, where mAPG and nmAPG indicate monotone and nonmonotone APG in (Li and Lin, 2015), and APG is the regular accelerated gradient descent without support projection. $\mathbf{y}$ is randomly chosen for 500 times and the average error of the 500 trials are reported. All the three optimization methods converge within 50 iterations for most cases. It can be observed that NAPGD-SP and MAPGD-SP converge faster than PGD and mAPG or nmAPG, demonstrating the empirical evidence of the provable convergence results in this paper.

Moreover, Figure 3 illustrates the results of monotone algorithms for the ℓ^0 regularized logistic regression, i.e. $\min_{\mathbf{x}} \frac{1}{n}\sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{x}' x_i)) + \lambda\|\mathbf{x}\|_0$ with $\lambda = 0.0001$ on the MNIST data set. The same experiment procedure is performed as that for the ℓ^0 penalized LSE problem. Note that the first part of the objective function is convex with bounded gradient.

It is worthwhile to mention that message-passing algorithms, such as approximate message-passing (AMP) (Donoho et al., 2009) and Expectation-Maximization AMP (EM-AMP) (Vila and Schniter, 2011), have also been used to solve optimization problems associated with compressive sensing. AMP and EM-AMP algorithms mainly handle compressing sensing problems where g is the squared loss function. On the other hand, our convergence results are established for general convex function g.



(a)



(b)

Figure 2: (a) Illustration of the objective value with respect to the iteration number for monotone algorithms (b) Illustration of the objective value with respect to the iteration for nonmonotone algorithms

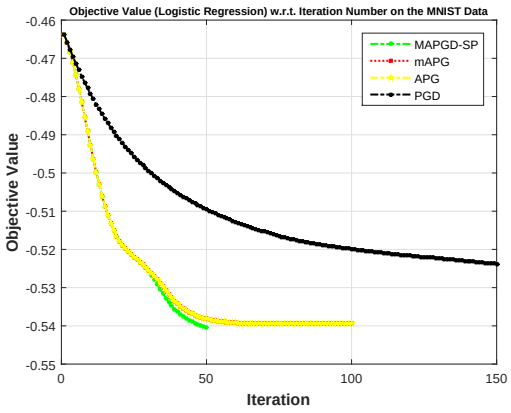


Figure 3: Illustration of the objective value with respect to the iteration number for ℓ^0 regularized logistic regression

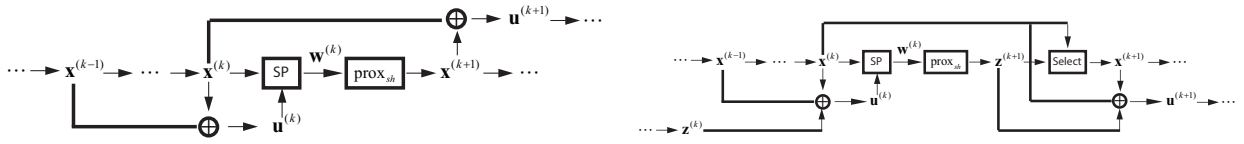
We also propose feed-forward neural networks as fast encoders to approximate the sparse codes generated by the proposed accelerated algorithms for ℓ^0 penalized LSE problem. We design an encoder, termed Deep-NMAPGD-SP, to approximate the optimization results of NMAPGD-SP. Each layer of Deep-NMAPGD-SP is designed to simulate the operations in each iteration of NMAPGD-SP. We use T pairs of data and the corresponding optimization results of NMAPGD-SP, i.e. $\{\mathbf{y}^{(t)}, \mathbf{x}^{(t)}\}_{t=1}^{T}$, as the training data for Deep-NMAPGD-SP. Deep-NMAPGD-SP is trained by minimizing the average ℓ^2 distance between the ground truth optimization results and the predicted results: $\frac{1}{N}\sum_{t=1}^{T} \|f(\mathbf{y}^{(t)}) - \mathbf{x}^{(t)}\|_2^2$. Inspired by (Gregor and LeCun, 2010), it is expected that the encoder designed in accordance with the optimization can approximate the optimization results by using far less number of layers than that of the original iterative optimization.

Similarly, a deep encoder named Deep-MAPGD-SP is designed to approximate the optimization results of MAPGD-SP. Figure 4 illustrates the architecture of Deep-NMAPGD-SP and Deep-MAPGD-SP. We using half of the optimization results on the MNIST data set as the training data for Deep-NMAPGD-SP, and the other half serve as the test data. The same setting is also appplied to Deep-MAPGD-SP. The test error of both Deep-NMAPGD-SP and Deep-MAPGD-SP are less than 0.007, demonstrating the effectiveness of fast encoders for the proposed algorithms in this paper.

We conduct an additional experiment with Deep-NMAPGD-SP and Deep-MAPGD-SP on the CIFAR-10 data set, which contains images of 10 categories with each category having 6000 images. We randomly choose 10000 images from the whole data set, and use half of the chosen images for training and the other half for test. The test error of both Deep-NMAPGD-SP and Deep-MAPGD-SP on the CIFAR-10 data set are 0.1792 and 0.2084 respectively. For experiments on the MNIST data set and the CIFAR-10 data set, the Deep-NMAPGD-SP and Deep-MAPGD-SP are trained for 300 epoches. The initial learning rate is 0.001, and it is divided by 10 at the 50-th epoch and the 100-the epoch during the training process.

8 APPENDIX

Proof of Theorem 5. It can be verified that when b is chosen according to (36), both $\mathbf{x}^*$ and $\bar{\mathbf{x}}^*$ are local solutions



(a) (b)

Figure 4: (a) Illustration of Deep-NMAPGD-SP for approximate NMAPGD-SP (b) Illustration of Deep-MAPGD-SP for approximate MAPGD-SP

to the capped-ℓ^1 problem as below:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2 + \mathbf{R}(\mathbf{x};b), \tag{38}$$

where $\mathbf{R}(\mathbf{x};b) = \sum_{j=1}^{n} R(\mathbf{x}_j;b)$, $R(t;b) = \lambda\frac{\min\{|t|,b\}}{b}$ for some $b>0$. It can be seen that $R(x;b)$ approaches the ℓ^0-norm when $b\to 0+$.

In the following text, let $\boldsymbol{\beta}_{\mathbf{I}}$ indicates a vector whose elements are those of $\boldsymbol{\beta}$ with indices in $\mathbf{I}$. Let $\Delta = \bar{\mathbf{x}}^* - \mathbf{x}^*$, $\tilde{\Delta} = \tilde{\partial}\mathbf{R}(\bar{\mathbf{x}}^*;b) - \tilde{\partial}\mathbf{R}(\mathbf{x}^*;b)$. We have

$$\|2\mathbf{D}^\top\mathbf{D}\Delta + \tilde{\Delta}\|_2 = 0.$$

It follows that

$$2\Delta^\top\mathbf{D}^\top\mathbf{D}\Delta + \Delta^\top\tilde{\Delta} \le \|\Delta\|_2\|2\mathbf{D}^\top\mathbf{D}\Delta+\tilde{\Delta}\|_2 = 0.$$

We now present another property on any nonconvex function P using the degree of nonconvexity defined as $\theta(t,\kappa) \triangleq \sup_s\{-\mathrm{sgn}(s-t)(\tilde{\partial}R(s;b)-\tilde{\partial}R(t;b))-\kappa|s-t|\}$ on the regularizer R. For any $s,t\in\mathbb{R}$, we have

$$-\mathrm{sgn}(s-t)\big(\tilde{\partial}R(s;b)-\tilde{\partial}R(t;b)\big) - \kappa|s-t| \le \theta(t,\kappa)$$

by the definition of θ. It follows that

$$\begin{aligned}\theta(t,\kappa)|s-t| &\ge -(s-t)\big(\tilde{\partial}R(s;b)-\tilde{\partial}R(t;b)\big)\\ &\quad -\kappa(s-t)^2 - (s-t)\big(\tilde{\partial}R(s;b)-\tilde{\partial}R(t;b)\big)\\ &\le \theta(t,\kappa)|s-t| + \kappa(s-t)^2.\end{aligned} \tag{39}$$

Applying (39) with $P=P_j$ for $j=1,\ldots,n$, we have

$$\begin{aligned}&2\Delta^\top\mathbf{D}^\top\mathbf{D}\Delta \le -\Delta^\top\tilde{\Delta} = -\Delta_{\mathbf{F}}^\top\tilde{\Delta}_{\mathbf{F}} - \Delta_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}^\top\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\\ &\le |\bar{\mathbf{x}}_{\mathbf{F}}^* - \mathbf{x}_{\mathbf{F}}^*|^\top\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa) + \kappa\|\bar{\mathbf{x}}_{\mathbf{F}}^* - \mathbf{x}_{\mathbf{F}}^*\|_2^2\\ &\quad + \|\Delta_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2\|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2\\ &\le \|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2\|\bar{\mathbf{x}}_{\mathbf{F}}^* - \mathbf{x}_{\mathbf{F}}^*\|_2 + \kappa\|\bar{\mathbf{x}}_{\mathbf{F}}^* - \mathbf{z}_{\mathbf{F}}^*\|_2^2\\ &\quad + \|\Delta\|_2\|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2\\ &\le \|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2.\end{aligned} \tag{40}$$

On the other hand, $\Delta^\top\mathbf{D}^\top\mathbf{D}\Delta \ge \kappa_0^2\|\Delta\|_2^2$. It follows from (40) that

$$2\kappa_0^2\|\Delta\|_2^2 \le \|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2\|\Delta\|_2 + \kappa\|\Delta\|_2^2 + \|\Delta\|_2\|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2.$$

When $\|\Delta\|_2\neq 0$, we have

$$\begin{aligned}&2\kappa_0^2\|\Delta\|_2 \le \|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2 + \kappa\|\Delta\|_2 + \|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2\\ &\Rightarrow \|\Delta\|_2 \le \frac{\|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2 + \|\tilde{\Delta}_{\mathbf{S}^*\cap\bar{\mathbf{S}}^*}\|_2}{2\kappa_0^2-\kappa}.\end{aligned} \tag{41}$$

According to the definition of θ, it can be verified that $\theta(t,\kappa) = \max\{0, \frac{\lambda}{b} - \kappa|t-b|\}$ for $|t|>b$, and $\theta(0,\kappa) = \max\{0,\frac{\lambda}{b}-\kappa b\}$. Therefore,

$$\begin{aligned}\|\theta(\mathbf{x}_{\mathbf{F}}^*,\kappa)\|_2 &= \Big(\sum_{j\in\mathbf{F}\cap\mathbf{S}^*}\big(\max\{0,\frac{\lambda}{b}-\kappa|\mathbf{x}_j^*-b|\}\big)^2\\ &\quad + \sum_{j\in\mathbf{F}\setminus\mathbf{S}^*}\big(\max\{0,\frac{\lambda}{b}-\kappa b\}\big)^2\Big)^{\frac{1}{2}}.\end{aligned} \tag{42}$$

In addition, for $k\in\mathbf{S}^*\cap\bar{\mathbf{S}}^*$, since $(\mathbf{D}^\top\mathbf{D}\Delta)_k = 0$ we have $\tilde{\Delta}_k = 0$. It follows that

$$\begin{aligned}\|\Delta\|_2 \le \frac{1}{2\kappa_0^2-\kappa}\Bigg(&\Big(\sum_{j\in\mathbf{F}\cap\mathbf{S}^*}\big(\max\{0,\frac{\lambda}{b}-\kappa|\mathbf{x}_j^*-b|\}\big)^2\\ &+ \sum_{j\in\mathbf{F}\setminus\mathbf{S}^*}\big(\max\{0,\frac{\lambda}{b}-\kappa b\}\big)^2\Big)^{\frac{1}{2}}\Bigg).\end{aligned} \tag{43}$$

This proves the result of this theorem. □

9 CONCLUSION

We present fast Proximal Gradient Descent (PGD) methods to solve the ℓ^0 regularization problem. We first prove improved convergence rate of PGD on the ℓ^0 regularization problem, then propose Nonmonotone Accelerated Proximal Gradient Descent with Support Projection (NAPGD-SP) and Monotone Accelerated Proximal Gradient Descent with Support Projection (MAPGD-SP) as fast algorithms for the ℓ^0 regularization problem. The potential of the proposed methods are evidenced by experiments.

References

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012. doi: 10.1214/12-AOS1032.

Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3858–3865, 2014.

A. Beck and M. Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11): 2419–2434, Nov 2009a.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Img. Sci.*, 2(1):183–202, March 2009b.

Thomas Blumensath and Mike E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5):629–654, 2008.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, August 2014. ISSN 0025-5610.

Kristian Bredies and Dirk A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*, 30(2):657–683, 2008.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45): 18914–18919, 2009.

M. Elad. Why simple shrinkage is still relevant for redundant representations? *IEEE Transactions on Information Theory*, 52(12):5559–5569, Dec 2006.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, March 2016a.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, Mar 2016b.

Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Z. Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages II–37–II–45. JMLR.org, 2013.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*, pages 399–406, 2010.

M. Hyder and K. Mahata. An approximate l0 norm minimization algorithm for compressed sensing. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 3365–3368, April 2009.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 685–693, 2014.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 379–387, 2015.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, March 2010.

L. Mancera and J. Portilla. L0-norm-based sparse representation through alternate projections. In *Image Processing, 2006 IEEE International Conference on*, pages 2089–2092, Oct 2006.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, May 2005.

Yu. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, Aug 2013.

Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM Journal on Imaging Sciences*, 8(1):331–372, 2015.

Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3115–3124, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

Joel A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

J. Vila and P. Schniter. Expectation-maximization bernoulli-gaussian approximate message passing. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 799–803, Nov 2011.

Cun-Hui Zhang and Tong Zhang. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statist. Sci.*, 27(4):576–593, 11 2012.